

Signaling Analysis for a Multi-Switch All-Optical Network

Gail C. Hudek and Douglas J. Muder

MITRE Corp.

Bedford, Massachusetts

Abstract

Three signaling schemes are analyzed for their potential implementation in a high-speed multi-switch all-optical network. This investigation is of interest because important characteristics of all-optical high-speed networks were not considered in the design of signaling protocols currently implemented in today's networks. Equations for throughput and delay are derived for the three schemes and comparisons are performed. Results obtained identify the signaling scheme to be implemented on an all-optical networking testbed.

1. Background and Introduction

In an all-optical network, the information-bearing signal remains in the optical domain from source to destination. Thus optoelectronic conversions are not performed. Some potential advantages of all-optical networks include: achievement of the terahertz bandwidth capacity of the fiber, improved error-rate performance and reliability, transport of any traffic type (including analog and digital).

Many challenges exist in implementing high-speed all-optical networks. This paper investigates the problem of network access and control, i.e., signaling. In the design of many existing signaling protocols, important characteristics of all-optical high-speed networks were not considered. These characteristics are: the propagation delay is comparable to or may exceed the transmission time; and optical buffers are not currently practical (queueing must take place at the hosts). In this paper, the performance of three signaling schemes is analyzed for an all-optical multi-switch network implementation. The results of this analysis are used to select a signaling scheme for implementation on the TBONE [1] (Testbed for All-Optical Networking) multi-switch all-optical network.

The network modeled is based on TBONE where all-optical crossbar switches transmit variable length bursts of data at gigabit speeds across MAN distances ($\sim 50km$). Each switch implements "burst-switching", where a connection is established through the switch only for the duration of a data burst. There are no collisions on this network but contention does exist; when several bursts arrive simultaneously at a switch and are destined for the same output port of that switch, one burst is selected. Out-of-band signaling channels are used where signaling packets are transported on channels separate from the data (see Figure 1).

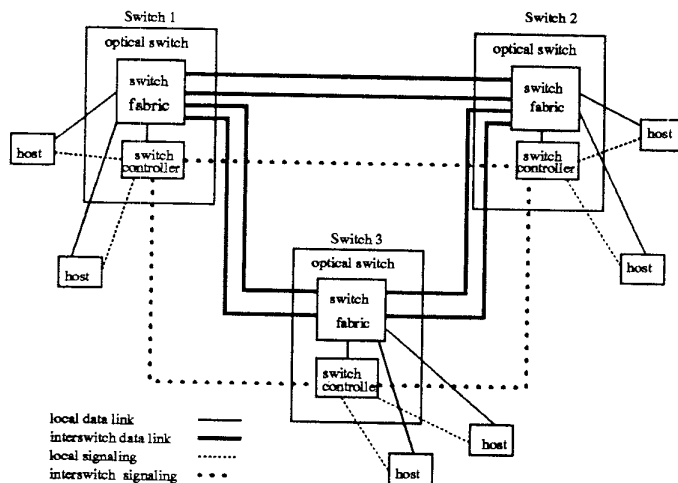


Figure 1: Three-switch all-optical network. Out-of-band signaling channels are used. $M = 3$, $N = 2$, $d = 2$, $m = 1$.

2. Overview of the Signaling Schemes

"Connect/Confirmation" (CC), "Tell-and-Go" (TAG), and "Reservation/scheduling with Just-in-Time switching" (RIT) signaling schemes are analyzed. CC is a simple scheme somewhat similar to that currently implemented in the telephone network and is analyzed for comparison. TAG and RIT schemes possess characteristics which help minimize the effect of propagation delay. In TAG, a host transmits a control packet and without waiting for confirmation from the network, transmits its data burst. Other tell-and-go schemes have been examined previously [2]; however, they assumed single-switch networks. In RIT, a host sends a request packet which is propagated to every switch. A scheduling agent residing in each switch then determines a schedule for all requests and notifies each host when to transmit its data burst. The connections in each switch are made just before the burst arrives. This RIT scheme assumes fixed shortest-path routing; in [3], a more complex algorithm performed routing and scheduling functions simultaneously. An analysis comparing signaling schemes similar to CC and TAG has recently been performed for an ATM network [4].

Synchronization is not required for the TAG or the CC schemes, but global synchronization of the switches is required for the RIT scheme. Since variable-size bursts are transmitted, disconnect packets are required in the CC and TAG schemes to notify the switches when a data burst is complete. In the RIT scheme, however, hosts include the duration of the data burst in the request packet so discon-

nect packets are not required.

Due to space limitations, in this paper only the final equations are given for each signaling scheme.

3. Assumptions and Definitions

In this analysis the throughput and delay are determined for each signaling scheme. The throughput (S) is defined as the average number of bursts successfully transmitted per burst transmission time. This S includes new plus retransmitted bursts. The “effective throughput” (S_{eff}), which includes only new arrivals and transmission attempts, is also determined. Delay (D) is defined to be the average time from when a burst is generated at a source until the entire burst is successfully received at the destination.

The network model used in the analysis consists of M identical switches. Each switch is directly connected to N hosts and d other switches (see Figure 1). All switch-to-host connections, referred to as local links (LLs), consist of a single pair of simplex fibers, one in each direction. Each LL is dedicated exclusively to one host. There are m pairs of simplex interswitch links (ISLs) connecting any two switches. The propagation delay is assumed to be the same across any LL, denoted τ_{LL} , and across any ISL, denoted τ_{ISL} . The diameter of the network (measured in ISLs) is denoted δ .

Every host is running multiple applications, each of which may attempt to transmit through the network. The aggregate application data burst generation process is modeled as a Poisson process at each host. The average generation rate of new bursts is λ bursts per second, and the burst length is exponentially distributed with average duration L seconds. The offered load is defined by $\alpha = \lambda L$. It is assumed that each time a burst is received at a switch, a new length is chosen independently from the exponential distribution [5]. With this independence assumption, the arrival of bursts at each link is also Poisson, with arrival rate λ_{ISL} at each ISL and λ_{LL} at each LL. The average length of time that an ISL is busy servicing a data burst is $1/\mu_{ISL}$, and for an LL this service time is $1/\mu_{LL}$.

It is assumed that each scheme (except TAG when no applications require retransmissions) continues trying to transmit a burst until it succeeds, i.e., there are no lost bursts. This means that at steady-state bursts must be successfully received at the same rate that new bursts arrive in the system. Therefore the effective throughput per host must be the same as the new burst arrival rate λ , and one can write,

$$S_{eff} = \lambda L.$$

Bursts which have not yet been successfully received queue at the host. Each burst is associated with a shortest path through the network (i.e., a fixed routing algorithm is assumed). Each path contains two LLs and $(N_s - 1)$ ISLs, where N_s is the number of switches in a path. The term h will also be used, where h is the *hop length* (in ISLs) of a

path.

It is assumed that all signaling packets (i.e., control packets, request packets¹ and disconnect packets) are transmitted on separate signaling channels and are always successfully received. The duration of a disconnect packet, denoted τ_{dp} , includes the processing and transmission time plus the time to release the connection in one switch. τ_s is the time to establish a connection through one switch. The switches along a given path are $\sigma_0, \dots, \sigma_h$. In CC and RIT, the probability that a newly generated burst has h hops in its path is γ_h .

4. Connect/Confirmation (CC)

In the CC scheme, a host transmits a request packet (containing the source and destination addresses) to the terminal switch σ_h . There the request waits in queue until the LL from σ_h to the receiving host is reserved. When that reservation is made, σ_h sends the request to σ_{h-1} , where it waits in queue to reserve one of the m ISLs from σ_{h-1} to σ_h . The reservation process proceeds backward along the burst’s transmission path until finally the LL from the transmitting host to σ_0 is reserved. The transmitting host is then notified and the data burst is sent.

The average waiting time in the burst request queue is denoted W_{BR} . By assumption, the h queues at the ISLs are identical, and their average waiting time is denoted W_{ISL} . The average waiting times for the transmitter and receiver LLs are W_{LLT} and W_{LLR} , respectively. Let τ_R be the time needed to transmit and process a request packet. The delay for an h -hop burst is,

$$D_h = L + 4\tau_{LL} + 3h\tau_{ISL} + 2(h+1)\tau_R + W_{BR} + W_{LLR} + W_{LLT} + hW_{ISL}.$$

5. Tell-and-Go (TAG)

In TAG, a host transmits a control packet (containing the source and destination addresses) and without waiting for confirmation from the network, transmits its data burst. At each switch in the network, the control packet is first received on the signaling channel, it is then processed and a connection through the switch is established (if available). Then immediately following, the data burst is received on the data link. It is assumed that fiber delay lines are used in each data path leading to a switch input port to account for the time to process the control packet and establish a connection through the switch.

The advantage of TAG over CC and RIT, is that hosts do not have to wait for confirmation or permission from the network before transmitting. In a high-speed network where the burst transmission time may be less than the propagation delay, this scheme could be quite effective in providing good performance. The disadvantage, of course,

¹The request packet also serves as a permission-to-transmit packet in the CC and RIT schemes.

is that hosts do not know before transmitting their data bursts, whether or not they will be blocked at any switch in the path. Therefore, Acks or Naks are required.

This analysis includes scenarios for both Acks (where the end host transmits an Ack after successfully receiving a data burst), and Naks (where the blocking switch transmits a Nak to the transmitting host). It is assumed that Acks and Naks will not experience contention and will always be successfully received.

For each of these Ack and Nak scenarios, comparative analyses are performed to determine the impact of:

1. ignoring the queue at the host transmitter. This queue resides in each host; its arrival rate, λ_i , is the aggregate arrival rate from all applications on the given host. (Many other analyses do not account for the delay incurred in this queue).
2. transmitting the control packet and data burst simultaneously vs. consecutively.
3. traffic requiring application-level retransmissions and traffic not requiring retransmissions (e.g., isochronous applications).

The following definitions and assumptions have not yet been discussed and are specific to the TAG scheme analysis.

- The arrival rate of newly generated plus *retransmitted* data bursts is assumed to be Poisson with average λ_s from each host.
- The arrival rate at each switch from an interswitch link is Poisson with average λ'_s .
- In the Nak scenario, a backoff algorithm is used for retransmission with exponentially distributed backoff time with average τ_b ; $\tau_b = L$ is assumed [6].
- The traffic distribution is defined by:
 - γ , the Pr [any burst from λ_s is destined for a local host], and
 - q , the Pr [an interswitch burst (i.e., from λ'_s) at a given switch is destined for other switches].
- λ_{ISL} is the aggregate average arrival rate to m output ISLs at any switch .
- λ_{LL} is the average arrival rate at any local link switch output port.
- P_B is the Pr [any switch in the path blocks the burst].
- P_{ISLB} is the Pr [a burst is blocked at an ISL].
- P_{ESB} is the Pr [a burst is blocked at an end switch in the path].

- τ_{cp} is the transmission time plus the processing time for one control packet for one switch or host.
- τ_{ack} (τ_{nak}) is the processing time plus the transmission time for one Ack (Nak) for one host or switch.
- ζ is the fraction of applications utilizing retransmissions.

To express throughput (S), for each type of link one can write,

$$S_{LL} = (1 - P_{ESB})\alpha_{LL},$$

$$S_{ISL} = (1 - P_{ISLB})\alpha_{ISL}.$$

The best-case and worst-case average delays for both the Ack and Nak scenarios were determined. The best-case corresponds to a data burst destined for a local host, and the worst-case corresponds to a data burst with N_s switches in its path. In this paper, only the worst-case average delays are presented. The queue at the host transmitter is modeled as an $M/M/1$ queue, so its waiting time is [5],

$$D_q = \frac{\lambda_i / \mu_{host}^2}{1 - \lambda_i / \mu_{host}}.$$

The worst-case delay is written for the Ack scenario,

$$D_{WCA} = \frac{\zeta P_B}{1 - P_B} [D_q + L + N_s(\tau_{cp} + \tau_{ss}) + 2(N_s - 1)\tau_{ISL} + 4\tau_{LL} + (\delta + 2)\tau_{ack}] + [D_q + L + N_s(\tau_{cp} + \tau_{ss}) + (N_s - 1)\tau_{ISL} + 2\tau_{LL}].$$

For the Nak scenario, allow j to represent the $(N_s - j)$ th switch which blocked the burst. Then the worst-case delay can be represented as follows,

$$D_{WCN} = \zeta \sum_{j=0}^{N_s-1} \frac{P_{LISLB}}{1 - P_{LISLB}} [\tau_b + D_q + (\tau_{cp} + \tau_{ss})(N_s - j) + 2\tau_{ISL}(N_s - (j + 1)) + 2\tau_{LL} + \tau_{nak}(N_s - j)] + [L + (\tau_{cp} + \tau_{ss})N_s + \tau_{ISL}(N_s - 1) + 2\tau_{LL} + D_q],$$

where,

$$P_{LISLB} = \begin{cases} (1 - P_{ISLB})^{N_s-1} P_{ESB} & j = 0 \\ (1 - P_{ISLB})^{N_s-(j+1)} P_{ISLB} & j > 0 \end{cases}$$

6. Reservation/scheduling with Just-in-Time Switching (RIT)

Both the CC and the TAG schemes have the potential to waste bandwidth: CC by reserving links longer than the burst duration, and TAG by transmitting some bursts several times. A reservation scheme attempts to use computing power and communication between the switches to avoid this potential waste of bandwidth.

W_{BR} , τ_R , and h are as defined in the CC analysis. Let τ_S be the time that the scheduler requires to check whether a given link is available at a given time. Since an h -hop burst requires the availability of $h + 2$ links, it is assumed that the scheduler takes $(h + 2)\tau_S$ to determine whether an h -hop burst can be sent at a given time.

The scheduling algorithm analyzed operates in parallel at each switch and attempts to schedule bursts in the order that they arrive. It will not attempt to improve efficiency by rescheduling previously scheduled bursts. The

transmitting host sends a request packet (containing the burst duration in addition to the source and destination addresses) to its local switch. The local switch time stamps the request and sends it to every other switch in the network. Each switch maintains a list of burst requests ordered by time stamp, and a record of the current schedule. The bursts whose time stamps are at least $\delta(\tau_{ISL} + \tau_R)$ old can be assumed to be on the list at every switch, and are referred to as *common* requests. The scheduler selects the common request with the oldest time stamp. The scheduler checks whether the first link (i.e., the transmitting host-to-switch LL) is available at time t , and whether the subsequent links are available at time t plus the appropriate propagation delay. If the answer is “yes”, the transmitting host is notified; if the answer is “no”, the question is considered again after a backoff period (exponentially distributed with average bL).

The delay of an h -hop burst is,

$$D_h = W_{BR} + 4\tau_{LL} + (\delta + h)\tau_{ISL} + (\delta + 3)\tau_R \\ + Q_h(W_S + (h + 2)\tau_S + bL) + (1 - b)L,$$

where, W_S is the average waiting time in the scheduling queue and Q_h is the number of times a request goes through the scheduling queue.

It is assumed that the scheduling queue has one server, Poisson arrivals, and an average service time of $h\tau_S$. So,

$$W_S = \frac{MN\lambda Q(h\tau_S)^2}{(1 - MN\lambda Qh\tau_S)}.$$

7. Results and Conclusions

The equations previously given for the three signaling schemes are valid for any symmetric network configuration. That is, each switch in the network has the same number of hosts and switches connected to it, and the same number of links interconnects each switch. Using these equations, the offered load is varied, and the throughput and delay are calculated for several different network sizes and configurations using a wide range of burst duration values. Other network parameter values (i.e., τ_{ss} , τ_{cp} , τ_R , τ_{dp} , τ_{ack} , τ_{nak}) used are believed to be representative of a high-speed network such as TBONE. In obtaining these results, it was assumed that half of all new bursts were destined for local hosts, and the destinations of the other half were uniformly distributed among nonlocal hosts (this assumption applies only to these particular curves but not to the overall analysis). The average worst case delays (i.e., for bursts with N_s switches in their paths) are shown in the following figures.

In general, it was determined that in addition to the offered load and the network configuration, a few important network parameter ratios dictate the relative throughput and delay performance of the schemes investigated.

7.1 TAG Results

The delay vs. throughput for TAG with $\zeta = 1$ (i.e., all traffic requires application-level retransmissions) is shown

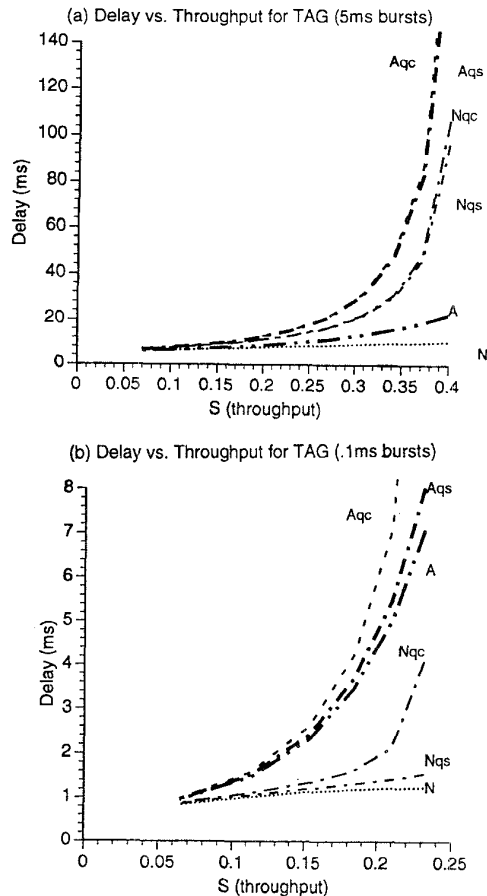


Figure 2: “A” refers to Ack scenario, “N” Nak scenario, “q” the queue at the host transmitter, “s” transmitting control packets and data bursts simultaneously, “c” transmitting them consecutively. The network configuration for both (a) and (b): $M = 10$, $N = 5$, $d = 3$, $\delta = 3$, $m = 2$, $N_s = 4$. In (a) $L/(\delta\tau_{ISL} + 2\tau_{LL})$ and L/τ_{cp} are large, in (b) both these ratios are small.

in Figure 2(a) and (b). Investigation of the different TAG scenarios with $\zeta = 1$ showed that the important ratios are: the burst duration to end-to-end propagation delay, $L/(\delta\tau_{ISL} + 2\tau_{LL})$, and L/τ_{cp} . When $L/(\delta\tau_{ISL} + 2\tau_{LL})$ is small, the Nak scenario shows much improvement over the Ack scenario (compare 2(a) vs. 2(b)). In addition, ignoring the queueing delay at the host transmitter leads to significantly erroneous results, particularly as the offered load increases.

When L/τ_{cp} is small, the TAG scheme performance of transmitting the control packet and data burst simultaneously has significant improvement over transmitting consecutively (see Figure 2(b)).

As one would expect, when $\zeta = 0$ (i.e., no traffic requires retransmissions) the throughput and delay performance are improved, and, of course, there is no performance difference between the Ack and Nak scenarios. In addition, the impact of ignoring the queue at the host transmitter is less significant. However, for these types of applications, the number of lost bursts is an important parameter and

must be considered. (The loss was determined but is not shown here.)

7.2 Comparison of Three Signaling Schemes

Since TAG is an inherently different signaling scheme than CC and RIT, slightly different approaches were used in deriving the delay for TAG than in CC and RIT. A parameter conversion (i.e., λ_s, γ, q for TAG; λ, γ_h for CC and RIT) was performed in order to compare the three signaling schemes (Details not shown here).

It is important to note that the throughput (S) in Figure 2 for TAG includes retransmissions. On the other hand, the effective throughput (S_{eff}), not including retransmissions, is represented in Figure 3 where the three signaling schemes are compared.

The optimum TAG scenario with $\zeta = 1$ (i.e., Nak scenario transmitting control packets and data bursts simultaneously) was used in comparing the three signaling schemes. Delay vs. throughput results for the three signaling schemes appear in Figure 3(a) and (b).

In comparing the CC, TAG, and RIT signaling schemes, the important ratios are: $L/(\delta\tau_{ISL} + 2\tau_{LL})$ and L/τ_S . When $L/(\delta\tau_{ISL} + 2\tau_{LL})$ is small, as anticipated for high-speed networks, TAG is much improved over CC (compare 3(a) vs. 3(b)). In addition, TAG is much improved over RIT when the offered load is low. As the offered load increases, the relative TAG and RIT performance depends on L/τ_S ($L/\tau_S = 50$ gives little improvement over TAG when $M=10, d=3$, as in 3(a) and (b)). Given the same value of L/τ_S , RIT gives less improvement over TAG as $L/(\delta\tau_{ISL} + 2\tau_{LL})$ grows smaller (e.g., as the network gets larger, or burst durations grow smaller for a given network).

In conclusion, although RIT shows some improvement over TAG under certain conditions, it would require a very fast processor ($\tau_S = .2\mu s$) and would not scale well. Therefore, these results indicate that TAG would be the most promising of these schemes to implement on an all-optical high-speed network where the burst transmission time may be less than the propagation delay. Hence, TAG will be implemented on the TBONE network.

Acknowledgement: This work was supported by ARPA under Air Force contract F19628-94-D-0001.

References

1. L. McAdams, I. Richer, and S. Zabele, "TBONE: TestBed for all-Optical Networking", IEEE/LEOS, July 1994.
2. S. F. Su, and R. Olshansky, "Performance of Multiple Access WDM Networks with Subcarrier Multiplexed Control Channels", J. Lightwave Tech., May/June 1993.

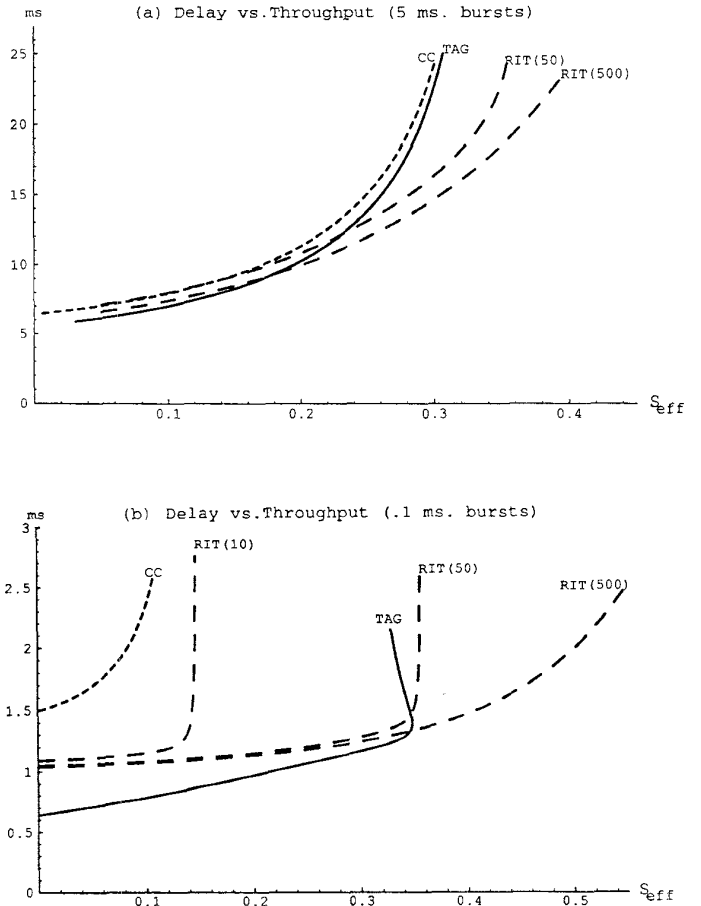


Figure 3: Network configuration for (a) and (b): $M = 10$, $N = 5$, $d = 3$, $\delta = 3$, $m = 2$, $N_s = 4$. RIT(x) refers to RIT with $x = L/\tau_S$. In (a) $L/(\delta\tau_{ISL} + 2\tau_{LL})$ is large and L/τ_S is as shown. In (b) $L/(\delta\tau_{ISL} + 2\tau_{LL})$ is small and L/τ_S is as shown.

3. C. G. Boncelet and D. L. Mills, "A Labeling Algorithm for Just-In-Time Scheduling in TDMA Networks", Univ. of Delaware, Jan. 1992.
4. A. Leon-Garcia, Personal communication, Univ. of Toronto, Canada.
5. L. Kleinrock, *Queueing Systems Vol. II*, J. Wiley and Sons, 1976.
6. L. Kleinrock and F. A. Tobagi, "Packet Switching in Radio Channels: Part I-Carrier Sense Multiple-Access Modes and Their Throughput-Delay Characteristics", IEEE Trans. Comm., Dec. 1975.